DOCUMENT RESUME

ED 357 068 TM 019 847

AUTHOR Klockars, Alan J.; Hancock, Gregory R.
TITLE An F-Enhanced Stagewise Protected Procedure for Testing a Complete Set of Planned Orthogonal Contrasts.
PUB DATE Apr 93
NOTE 7p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Comparative Analysis; *Computer Simulation; *Hypothesis Testing; Mathematical Models; Monte Carlo Methods; Probability
IDENTIFIERS F Test; *Multiple Comparisons; Null Hypothesis; Omnibus Tests; Orthogonal Comparison; Power (Statistics); *Stagewise Protected Procedure; Type I Errors

ABSTRACT

The challenge of multiple comparisons is to maximize the power for answering specific research questions, while still maintaining control over the rate of Type I error. Several multiple comparison procedures have been suggested to meet this challenge. The stagewise protected procedure (SPP) of A. J. Klockars and G. R. Hancock tests null hypotheses sequentially, with each stage of testing preceded by an omnibus F-test on the remaining between-group variability. A modification of this SPP is proposed and tested through a Monte Carlo simulation. A redirection of the Type I error rate in protected tests is proposed to be an indirect reference to the probability that both the omnibus and contrast null hypotheses will be rejected. Simulation results indicate that the proposed adjustment of the alpha-level for the stagewise omnibus F-tests appears to provide the needed power to the original SPP where it was previously weak, and does so without increasing the Type I error rate for the experiment beyond the nominal alpha-level. Two tables present analysis results. (SLD)

An F-enhanced stagewise protected procedure for testing a complete set of planned orthogonal contrasts

Alan J. Klockars, University of Washington
Gregory R. Hancock, Auburn University

Background

The challenge of multiple comparisons is to maximize the power for answering specific research questions while still maintaining control over the rate of Type-I error. In the last 14 years a number of modified multiple comparison procedures have been suggested to answer this challenge. Many of these involve the sequential testing of null hypotheses ordered according to the tenability (i.e., p-values) of their null hypotheses (see Holm,1979). A method offered by Hochberg (1988) proceeds in a "step-up" manner, sequentially retaining null hypotheses from the most to the least tenable. Conversely, procedures by Holm (1979) and Shaffer (1986) are "step-down" in that they seek to reject null hypotheses starting from the least tenable. Klockars and Hancock (1992) proposed a modification of Shaffer's method, specifically applicable to a complete set of planned orthogonal contrasts. Their "stagewise protected procedure" (SPP) tested null hypotheses sequentially, with each stage of testing preceded by an omnibus F-test on the remaining between-group variability. The current paper proposes a modification of this SPP.

The SPP offered potential gains in power through the logical implications of the omnibus F-test (conducted at α=.05) at each stage. Specifically, rejection of an initial omnibus null hypothesis logically implied that at most j-1 of the j orthogonal contrast null hypotheses could be true. Using Bonferroni's inequality, the per comparison (per contrast) error rate pursuant to a rejected omnibus null hypothesis was $\alpha/(j-1)$ rather than α/j. If the test statistic for the contrast with the smallest p-value exceeded the corresponding critical value, the experimenter proceeded to the next omnibus test. In this stage the experimenter constructed an F-ratio for a partial omnibus test, the numerator of which consisted of the pooled treatment variance and degrees of freedom from the remaining j-1 orthogonal contrasts. If this partial omnibus test rejected its null hypothesis, the experimenter concluded that at most j-2 of the remaining j-1 contrast null hypotheses could be true. The appropriate per contrast error rate was then $\alpha/(j-2)$. If the contrast with the second smallest p-value was significant at the $\alpha/(j-2)$ level, the process continued. In a similar manner the variances associated with all remaining contrasts were pooled and the partial omnibus test evaluated at the .05 level. If judged significant the largest remaining contrast was evaluated against a reduced critical value, $\alpha/(j-i)$ at the ith stage of testing.

In their simulation of the SPP, Klockars and Hancock showed the method offered more power than those of Hochberg and Shaffer when there was much overall variation in the groups. That is, if all four contrasts in the simulation were altered by treatment effects, the SPP provided considerably more power than other methods simulated -- under all configurations of small,

medium and large treatment effects. However, when only one or two of the four contrasts were altered by treatment effects, the SPP provided considerably less power than other methods simulated. This lack of power of the SPP with few treatment effects was primarily due to a conservative level of control over Type-I error. In order for a replication to be judged significant it nad to reject both the omnibus and the contrast null hypotheses. The effective Type-I error rate for the least tenable contrast when the overall omnibus null hypothesis was true (and thus the contrast null hypothesis was also true) was approximately .035 rather than the nominal .05 level. Similarly, the Type-I error rate when some (but not all) of the contrast null hypotheses were true continued to be smaller than the nominal .05 (though the magnitude of the difference was not as great as with the overall omnibus null hypothesis).

As is the purpose of protected tests, the maximum possible value of the experimentwise Type-I error rate is determined by the α level of the omnibus test. If α =.05 for the omnibus test, only 5% of all experiments will produce an omnibus F value (F_O) sufficient to reject the omnibus null hypothesis when the overall null hypothesis is true. The tests of the largest contrast are only conducted within that subset of the experiments that rejected the omnibus null hypothesis. The presence of a significant omnibus test does not guarantee that there will be a significant contrast. When this occurs the total proportion of contrast null hypotheses resulting in Type-I error will be less than the stated α level of 5%. It is the purpose of this paper to propose a redefinition of the Type-I error rate in protected tests to be in direct reference to the probability that both the omnibus and contrast null hypotheses will be rejected. That is, it is proposed herein that the nominal α refer to the probability of the joint occurrence of the observed omnibus test statistic (F_O^*) exceeding a stated critical value (F_O) and the contrast test statistic (F_C^*) exceeding the required critical value (F_C). Symbolically, then, given that the overall omnibus null hypothesis is true --

$$\alpha = Pr[(F_O^* > F_O) \cap (F_C^* > F_C)].$$

Setting F_O and F_C according to the above criterion allows the experimenter to set the critical values of the omnibus and contrast tests at α levels greater than the nominal value as long as the joint probability is maintained at α. There is a continuum of paired (F_O, F_C) values that would provide an overall Type-I error rate of α. To eliminate this indeterminacy, the value of F_C is herein defined as in Klockars and Hancock (1992). That is, F_C for the largest contrast is set at the α-level for the first contrast conducted when the overall omnibus null hypothesis is rejected -- $\alpha/(j-1)$, or .05/(4-1)=.0167 for the case of four orthogonal contrasts. (This would translate to a maximum possible experimentwise error rate of (4)(.0167)=.0667 if all four contrast null hypotheses were true; but this would be logically inconsistent with the rejected overall omnibus null hypothesis.) Tests of

additional contrasts within the sequence would be conducted with α similarly defined for the reduced number of true null hypotheses possible.

Method

The adjusted α-levels and corresponding F_O for the omnibus and partial omnibus tests were obtained by a Monte Carlo simulation. Using the previously defined values of F_C, the corresponding F_O values were found such that the joint probability of the omnibus test statistic F_O* exceeding F_O and the largest contrast test statistic F_C* exceeding F_C was .05. With five treatment groups, values were needed for the overall omnibus test with four contrasts, and for partial omnibus tests based on three and two contrasts. The effective α-levels for the omnibus and partial omnibus tests were found to be .080 for four contrasts, .064 for three contrasts, and .052 for two contrasts.

The effect of this modified definition of α on the experimental power was evaluated with a separate Monte Carlo simulation in which this F-enhanced SPP was compared to the original SPP, to Shaffer's sequentially rejective Bonferroni procedure, and to a standard Bonferroni procedure (for reference). Scores for five groups of n=16 observations were generated in a FORTRAN program using the Box and Muller (1958) transformation to convert randomly drawn pairs from a uniform distribution into random normal deviates. The treatment sum of squares was partitioned into the four orthogonal contrasts defined by the Helmert series. Treatment effects were created by the appropriate addition of constants onto the generated scores. A small effect was defined as one that produced a contrast that was correctly detected by the standard Bonferroni procedure 20% of the time. Medium and large effects were similarly defined by their power in the standard Bonferroni procedure being 50% and 80%, respectively.

Four patterns of treatment effects were simulated, representing varying amounts of between-group variance. The first had only one small effect and the remaining three contrasts had true null hypotheses (NNNS). The second had a small and a medium effect with two true null hypotheses (NNSM). The third had a small, medium, and large effect along with one true null hypothesis (NSML). The fourth had a small, two medium, and one large effect (SMML). In addition, the case of four true null hypotheses (NNNN) was simulated. Twenty-thousand replications were run for each treatment effect configuration.

Results and Discussion

The power of all methods compared under the various simulated treatment effect configurations are presented in Table 1. The Bonferroni power figures reflect the definitions of the three magnitudes of treatment effects. The pattern of results for Shaffer and the original SPP replicates the findings in Klockars and Hancock (1992). The power for detecting treatment effects

in both the NNNS and NNSM configurations is lower for the original SPP than Shaffer's method, while the reverse is found as more treatment variability is introduced (NSML and SMML). The new finding offered by this paper involves the F-enhanced SPP. Specifically, it has essentially equal or greater power than Shaffer's method under all treatment effect conditions. Where only one small effect is present (NNNS), the F-enhanced SPP is only .005 lower than Shaffer. Where there are more treatment effects the F-enhanced SPP offers up to 4% more power than Shaffer, and has slightly more power than the original SPP (although this superiority averages less than 1%).

Insert Table 1 about here

The per experiment Type-I error rates (i.e., the sum of per contrast error rates) for the simulations are presented in Table 2. These values represent an upper limit for the experimentwise error rate. As expected the error rates for the F-enhanced SPP are closer to .05. The increased power of the method, particularly for detecting smaller effects, is accomplished by more closely matching the observed error rate with the nominal rate.

Insert Table 2 about here

Conclusion

The proposed adjustment of the α-level for the stagewise omnibus F-tests appears to provide needed power to the original SPP in those situations where it previously was weak, and does so without increasing the Type-I error rate for the experiment beyond the nominal α-level. This simulation is limited, of course, as it was undertaken only for a particular number of treatment groups, a particular sample size, and a particular α-level. Further research should be conducted to assess the impact of the modified SPP in a variety of situations.[1]

1 A first attempt at constructing a table of required F_0 values using a Monte Carlo approach is available from the authors.

Table 1

Power ratings of multiple comparison procedures

Configuration	Effect	Bonferroni	Shaffer	Original SPP	F-enhanced SPP
NNNS	Small	.189	.193	.154	.188
NNSM	Small	.196	.215	.207	.224
	Medium	.494	.507	.481	.516
NSML	Small	.196	.251	.264	.271
	Medium	.496	.546	.556	.572
	Large	.800	.822	.828	.840
SMML	Small	.193	.321	.360	.361
	Medium*	.496	.612	.650	.653
	Large	.799	.854	.878	.880

(*average across two contrasts)

Table 2

Per experiment Type-I error rates

Configuration	Bonferroni	Shaffer	Original SPP	F-enhanced SPP
NNNN	.049	.050	.036	.050
NNNS	.040	.044	.042	.050
NNSM	.025	.033	.035	.042
NSML	.013	.026	.035	.035

References

Box, G. E. P., & Muller, M. E. (1958). A note on the generation of random normal deviates. Annals of Mathematical Statistics, 28, 610-611.

Hochberg, Y. (1988). A sharper procedure for multiple tests of significance. Biometrika, 75, 800-802.

Holm. S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6, 65-70.

Klockars, A. J., & Hancock, G. R. (1992). Power of recent multiple comparison procedures as applied to a complete set of planned orthogonal contrasts. Psychological Bulletin, 111, 505-510.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81, 826-831.